The Comparability of Recommender System Evaluations and Characteristics of Docear's Users

Stefan Langer
Docear
Magdeburg
Germany
langer@docear.org

Joeran Beel
Docear
Magdeburg
Germany
beel@docear.org

ABSTRACT
Recommender systems are used in many fields, and many ideas have been proposed how to recommend useful items. In previous research, we showed that the effectiveness of recommendation approaches could vary depending not only on the domain of a recommender system, but also on the users' demographics. For instance, we found that older users tend to have higher click-through rates than younger users. This paper serves two purposes. First, it shows that reporting demographic and usage-based data is crucial in order to create meaningful evaluations. Second, it reports demographic and usage-based data on Docear's recommender system. This sets our previous evaluations into context and helps others to compare their results with ours.

Categories and Subject Descriptors
H.3.4 [**Information Systems**]: Systems and Software – *performance evaluation (efficiency and effectiveness), user profiles and alert services.*

General Terms
Management, Measurement, Performance, Effectiveness, Human Factors

Keywords
Recommender system, Evaluation, Demographic Information, User Characteristics

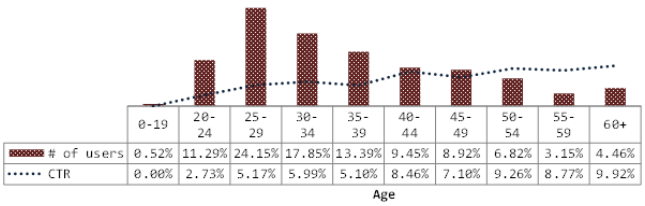
1. INTRODUCTION
Research paper recommender systems aid researchers in finding relevant literature. This includes recommending newly published articles in the researcher's field of interest, or recommending serendipitous literature of neighboring topics that researchers might not find through selective search.

More than 80 approaches to research paper recommender systems were published between 1998 and 2012 in more than 170 articles [5]. A thorough evaluation is crucial to identify individual strengths and weaknesses of the approaches, or to identify maybe even *the* most effective approach.

Evaluations should provide others with the knowledge necessary to identify the most promising approaches for a given task. In order to create replicable results of evaluations the *first step* is to identify the factors, which influence a recommender system's performance. As a *second step* data on these factors have to be published.

Weber and Costillo found that a typical female US web user thinks about the composer Richard Wagner when searching for the term "wagner", while typical male US users are more likely to be referring to the paint sprayer company Wagner [15]. This example illustrates how different the usefulness of a given recommendation or search result may be judged by user groups with different demographics. Krulwich was first to describe how demographic clusters of user profiles can be created and that they are useful in order to improve online information targeting, such as web advertising [11]. Demographic data influence the response to email marketing [9]. It can also be used to recommend restaurants [12], music [14], or movies [13] to users. All of these papers show that demographic data can be used to find suitable search items for users. However, we are not aware of any discussion about which user characteristics should be reported in research papers, and what the impact of these characteristics is on a recommender system's effectiveness.

There are many papers on the evaluation of recommender systems regarding comparability of results. Herlocker et al., for instance, stress the necessity of publishing information like the user task and properties of the data sets being used in an evaluation (e.g. domain, inherent and sample features) [10]. They also state the importance of describing the way in which the prediction quality is measured and that a user's satisfaction with a recommender system not only depends on its accuracy. While they discuss user-based evaluations of a recommender system, they mainly focus on differences in how the evaluation (e.g. laboratory studies vs. field studies or explicit vs. implicit ratings) is conducted. They neglect effects of the user's demographics or behavior on the effectiveness of recommender systems.



	0-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60+
# of users	0.52%	11.29%	24.15%	17.85%	13.39%	9.45%	8.92%	6.82%	3.15%	4.46%
CTR	0.00%	2.73%	5.17%	5.99%	5.10%	8.46%	7.10%	9.26%	8.77%	9.92%

Age

Figure 1: CTR and User Distribution by Age [7]

In a recent study, we found demographic data to influence the effectiveness of recommender systems in general [7]. In that study, older people had higher *Click-Through-Rates* (CTR) than younger users (Figure 1). Using CTR as measure for effectiveness considers clicks as positive implicit ratings and measures *precision* of the recommendation approach. If, for instance, a user clicked five out of 100 recommendations, CTR is 5%.

In this paper, we pursue two research objectives.

First, we explore the impact of user characteristics on the effectiveness of recommendation approaches. To accomplish this objective, we analyze whether identical recommendation approaches perform differently for different user groups, and which user characteristics are responsible for the differences. Based on our findings, we propose that future evaluations of recommender systems

should report their user characteristics in order to make their evaluations more meaningful and comparable.

Second, provide information about the users of Docear's research paper recommender system. This should help researchers to better understand our previous research [1, 4, 8], since our previous evaluations did not provide detailed information about Docear's users.

2. METHODOLOGY

We conducted our research based on Docear, which is an open-source reference management software that helps researchers in their daily work with academic literature. Docear stores information like the papers users read or which information of their papers they highlighted. Docear stores information in mind maps [2]. Users can augment these mind maps with own ideas, create categories to organize their annotations and finally use Docear to draft their own papers or research thesis. Apart from text, Docear mind maps can contain pictures, web links, file links, LaTeX formulas, comments, rich formatted text or citation information (Figure 2).

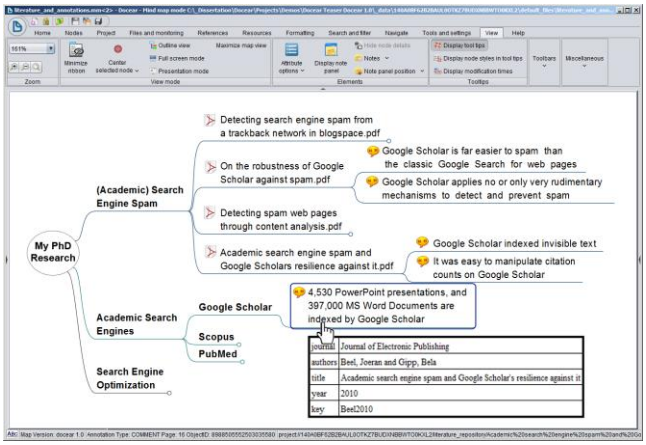


Figure 2: Screenshot of a mind map in Docear

Docear offers a research paper recommender system, which is activated by default. Users can deactivate it during Docear's installation process and from Docear's preferences. Users with enabled recommendations settings *automatically* receive pre-generated recommendation sets of up to 10 documents every 5 days (Figure 3). Users can also *request* recommendation manually.

To generate recommendations, Docear uploads a user's mind maps to the recommender system, where the mind-maps are stored in a graph database. Older versions (revisions) of the same mind map are stored as backup for the user and may eventually be used in order to determine *concept drift*. The recommender system creates algorithms based on randomly selected parameters and generates models of the user's interest [3, 6]. To find relevant articles, the recommender system compares these *user models* to our digital library, which currently contains around 1.8 million academic articles in full text.

To achieve our first research objective, we implemented different variations of Docear's recommendation approaches. We then analyzed the effectiveness of the variations for different groups of users. As a measure to compare the influence of specific demographic or usage-based factors on the effectiveness of Docear's recommender system, we use click-through rate (CTR).

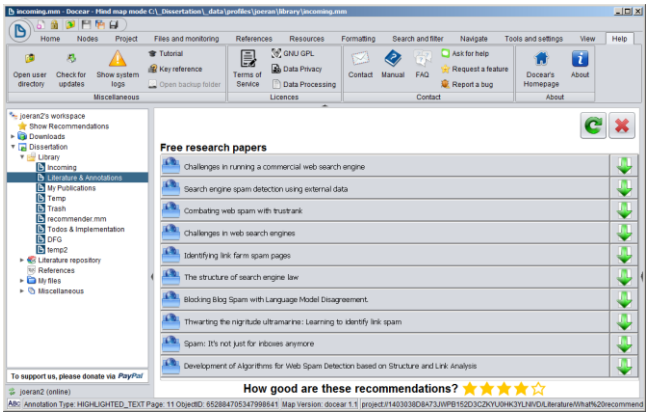


Figure 3: Recommendations in Docear

To achieve our second research objective, we provide information on the user's demographics, i.e., ages and genders, and their distributions among Docear's users. We also provide usage-based information e.g., the number of mind maps, mind map nodes and linked research papers.

It is important to note that we collected demographic data of our users in different ways. If users register with Docear using the project's website[1], they can optionally provide information on their gender and year of birth. We use these data to evaluate the effectiveness of our recommender system. Data on the usage of our website or the introduction video is based on Google Analytics and YouTube.

For our research, we analyzed 240,948 recommendations in 25,355 recommendation sets, recommended to 4,164 Docear users between April 2013 and June 2014 (Figure 4). The majority (76.1%) of the sets was delivered automatically. Users explicitly requested recommendations 23.1% of the times.

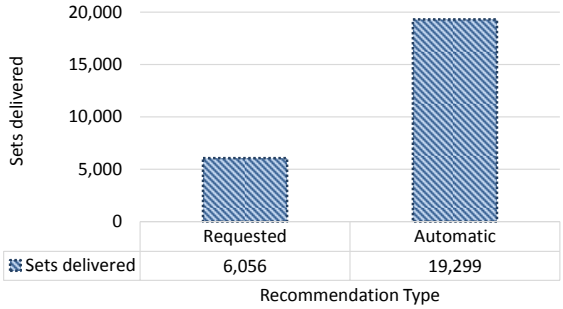
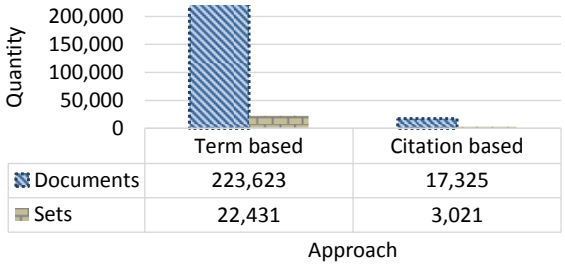


Figure 4: Number of delivered sets for requested and automatic recommendations

In our analysis, we distinguish between the CTR that the *term* based and *citation* based approaches achieved. Both approaches use *Content based Filtering* (CBF) either on the terms or citations (links to papers) that the users' mind maps contain. Since the coverage of our digital library is limited, the users' citations were not always identified. As a result, the citation based approach often failed to recommend any documents to the user. Our observations are hence

[1]http://docear.org

based primarily on term based recommendations (Figure 5), and the significance of citation based data should be considered with caution.



	Term based	Citation based
Documents	223,623	17,325
Sets	22,431	3,021

Figure 5: Quantity of Term Based vs. Citation Based data

3. RESULTS

Section 3.1 shows how different user groups (distinguished by gender and age) use Docear and associated resources or services. Section 3.2 presents how user characteristics, like age, gender or the user type, influence the CTR achieved by Docear's recommender system. Section 3.3 focuses on how the amount of a user's data influences the CTR of the recommender system. Apart from discussing the amount of a user's mind maps and revisions, it also studies some other usage-based factors.

3.1 System Usage by Demographics

We observed that males and females diverge in the general usage of Docear. The fraction of male users increases the more intensive Docear or associated resources are used (Figure 6). The majority of Docear's website visitors are male (69.16%) and a larger proportion of the users who watch Docear's introductory video are male (78%). Most registered users are males (80.62%), and of the users who used Docear on at least seven or more days an even larger fraction is male (85.62%). Apparently, the concept of Docear is more attractive for males than for females. It would be interesting to study whether the same is true for other reference management or mind mapping tools.

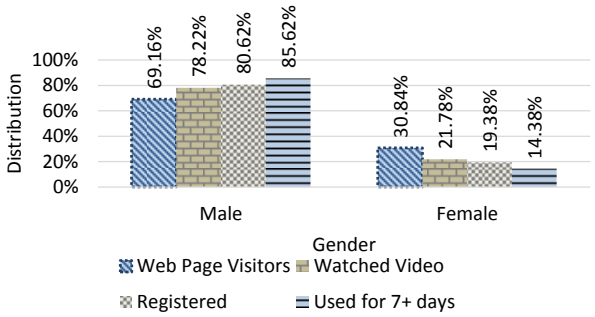
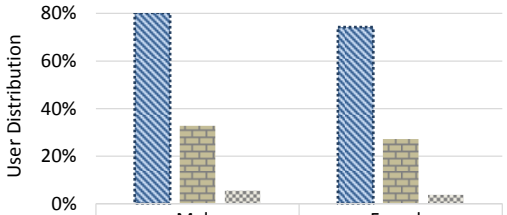


Figure 6: System Usage by Gender

While 80.21% of Docear's male users activated the recommender system, only 74.3% of the female users did (Figure 7). Of the male users, 32.75% received at least one set of recommendations, while only 27.19% of female users did. On the other hand, only 3.76% of female users, but 5.52% of male users deactivated the recommender system after receiving at least one set of recommendations. In total, 1,032 male users received recommendations, while only 186 female users did.

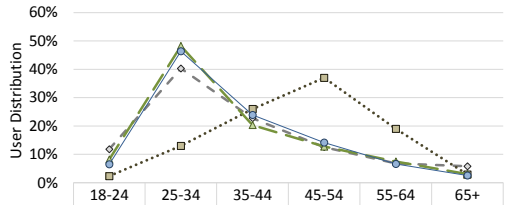
The usage of Docear and its recommender system does not only differ by gender but also by the users' age.

The group of users aged from 25 to 34 represents the group with the most website visitors (40.29%), registered users (48.23%) and users who use Docear for seven and more days (46.34%) (Figure 8). The second largest group using Docear for at least a week is between 35 and 44 years old. Very young users (18 to 24 years of age) and older users (55+ years of age) make only a small fraction of Docear's user base. It stands out that the proportions of user groups watching Docear's introductory video differs significantly from the other interaction types. While only 12.76% of the registered users are between 45 and 54 years old, 37% of the users watching the introductory video are between 45 and 54 years old. In contrast, only 13% of the users watching the video are between 25 and 34 years old, although they represent 40% of the website visitors and 48% of the registered users.



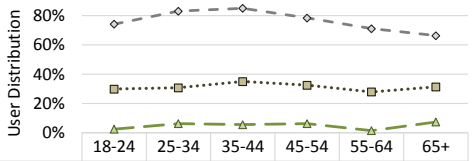
	Male	Female
Recs active	80.21%	74.30%
Received 1+ sets	32.75%	27.19%
Recs deactivated	5.52%	3.76%

Figure 7: Recommendation Usage by Gender



	18-24	25-34	35-44	45-54	55-64	65+
Web Page Visitors	11.81%	40.29%	22.76%	12.53%	6.81%	5.80%
Watched Video	2.30%	13.00%	26.00%	37.00%	19.00%	2.90%
Registered	8.19%	48.23%	20.32%	12.76%	7.47%	3.04%
Used for 7+ days	6.50%	46.34%	23.88%	14.13%	6.61%	2.54%

Figure 8: System Usage by Age



	18-24	25-34	35-44	45-54	55-64	65+
Recs active	74.20%	83.09%	85.01%	78.40%	71.11%	66.34%
Received 1+ sets	29.85%	30.76%	35.04%	32.50%	27.92%	31.40%
Recs deactivated	2.50%	6.27%	5.62%	6.29%	1.49%	7.41%

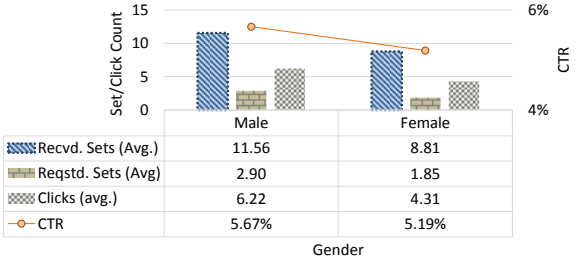
Figure 9: Recommendation Usage by Age

Regarding Docear's recommender system, users between 35 and 44 years of age seem to be interested most in recommendations (Figure 9). A relatively high percentage of this group (85.01%) activated the recommender system and received at least one set of recommendations (35.04%). Users older than 65 years seem to be least interested in recommendations. Only 66.34% of these users activated the recommender system and a relatively large percentage (7.41%) deactivated it after receiving at least one set of recommendations. While very young users between 18 and 24 years and relatively old users between 55 and 64 are least likely to deactivate the recommendations (2.5% and 1.49% respectively) later,

they are also not very likely to activate them in the first place (36.31% and 35.66% respectively).

3.2 Recommendation Effectiveness by User Characteristics

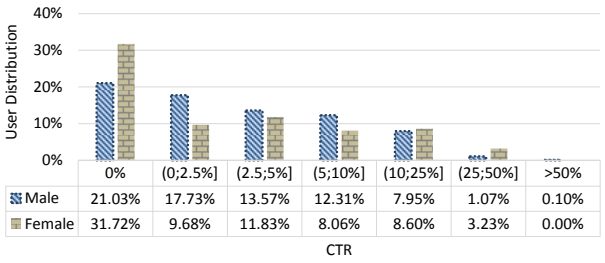
On average, male users have higher CTR (5.67%) than female users (5.19%). They also receive more recommendation sets, request recommendations more often, and click recommendations more often (Figure 10).



	Male	Female
Recvd. Sets (Avg.)	11.56	8.81
Reqstd. Sets (Avg)	2.90	1.85
Clicks (avg.)	6.22	4.31
CTR	5.67%	5.19%

Figure 10: Averages and CTR by Gender

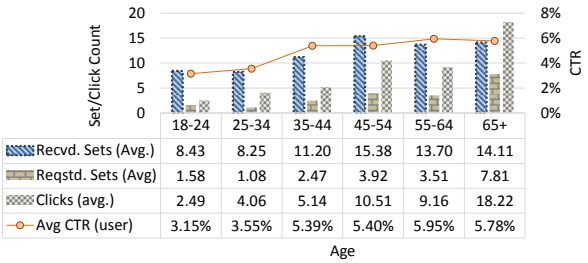
Figure 11 shows that around a third (31.72%) of the female users never clicked any recommendation (CTR of 0%). In contrast, only a fifth (21.03%) of male users never clicked recommendations. About a tenth of the users (11.83% of the females and 9.12% of the males) seemed to be strongly interested in the recommendations with a CTR of 10% and more. Small fractions of both male (1.17%) and female users (3.23%) have CTR above 25%.



	0%	(0;2.5%]	(2.5;5%]	(5;10%]	(10;25%]	(25;50%]	>50%
Male	21.03%	17.73%	13.57%	12.31%	7.95%	1.07%	0.10%
Female	31.72%	9.68%	11.83%	8.06%	8.60%	3.23%	0.00%

Figure 11: CTR Ranges by Gender

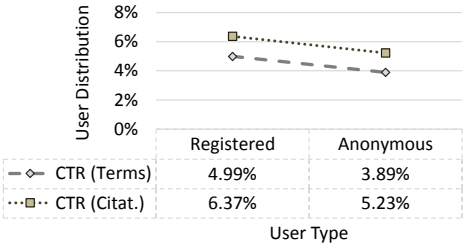
CTR of Docear's recommender system also differs by age (Figure 12). Users being 34 years and younger have CTR of around 3% to 3.6% on average, while older users have higher average CTR between 5% and 6%. This is particularly interesting, since older users activated recommendations less often, and hence seemed less interested in recommendations. Apparently, those older users who decide to activate recommendations are more interested in the recommended documents than younger users, who tend to activate recommendations more often while having lower CTR on average. It may also indicate that among users, who are not interested in recommendations, older users deactivate the recommender system, while younger users simply ignore automatically received recommendations. Since activating Docear's recommender system means that the user's mind maps are automatically transferred to the recommender system, it may also hint that older users are generally more concerned about data privacy than younger users.



	18-24	25-34	35-44	45-54	55-64	65+
Recvd. Sets (Avg.)	8.43	8.25	11.20	15.38	13.70	14.11
Reqstd. Sets (Avg)	1.58	1.08	2.47	3.92	3.51	7.81
Clicks (avg.)	2.49	4.06	5.14	10.51	9.16	18.22
Avg CTR (user)	3.15%	3.55%	5.39%	5.40%	5.95%	5.78%

Figure 12: Averages and CTR by Age

Docear has two different kinds of users, *registered* and *anonymous* users[2]. While registered users registered themselves either in Docear or on Docear's websites, anonymous users are only recognized by an anonymous user token. On average, anonymous users have lower CTR for term-based recommendations (3.89%) than registered users (4.99%) (Figure 13). Anonymous users also have lower average CTR for citation-based recommendations (5.23%) than registered users (6.37%). We suspect the attitude of registered users towards Docear and its recommender system to be slightly more positive. Anyway, for both registered and anonymous users, citation-based recommendations resulted in higher CTR than term-based approaches.

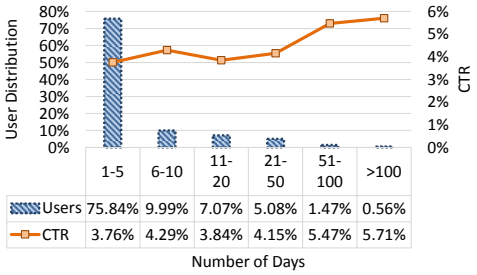


	Registered	Anonymous
CTR (Terms)	4.99%	3.89%
CTR (Citat.)	6.37%	5.23%

Figure 13: CTR by User Type

3.3 Recommendation Effectiveness by Usage

Apart from demographic data, there are other, usage based, factors influencing the effectiveness of Docear's recommender system. There is a tendency that the more intensively a user uses Docear, the higher CTR becomes. We observed this correlation in the following situations.



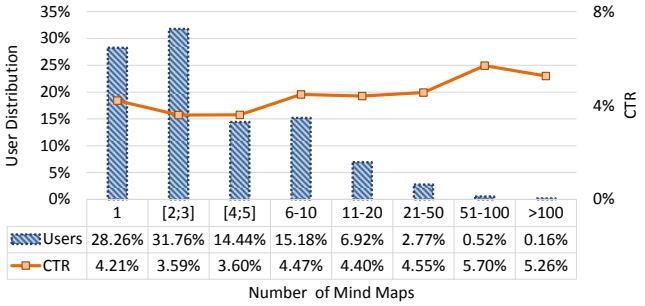
	1-5	6-10	11-20	21-50	51-100	>100
Users	75.84%	9.99%	7.07%	5.08%	1.47%	0.56%
CTR	3.76%	4.29%	3.84%	4.15%	5.47%	5.71%

Figure 14: Number of Days of Docear being used

The more often users start Docear, the higher their average CTR (Figure 14). Users, who started Docear on one to five days, have an average CTR of 3.76%. Users who start Docear on more than 100 days have an average CTR of 5.71%. For users who used Docear

[2] Docear can also be used as a „local" user without access to Docear's online features, but this type of user is of no relevance for this paper.

between 11 and 50 days, average CTR is lower and does not follow the trend. Unfortunately, for the overall effectiveness, the majority (75.84%) of Docear's users are those who started Docear only between one and five days, and have a low CTR on average. Only a small fraction (0.56%) of users started Docear on more than 100 days, but they have the highest CTR of 5.71% on average.

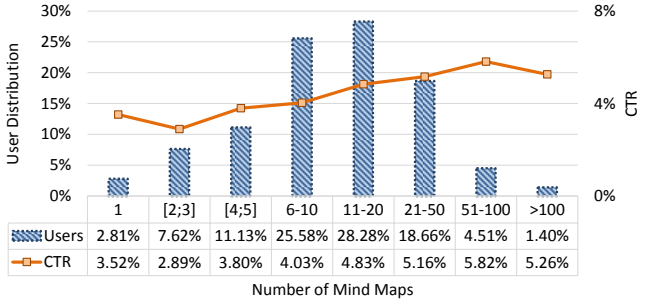


	1	[2;3]	[4;5]	6-10	11-20	21-50	51-100	>100
Users	28.26%	31.76%	14.44%	15.18%	6.92%	2.77%	0.52%	0.16%
CTR	4.21%	3.59%	3.60%	4.47%	4.40%	4.55%	5.70%	5.26%

Number of Mind Maps

Figure 15: Number of Mind-Maps (All Users)

There is a tendency that CTR increases the more mind maps users created (Figure 15). Average CTR for users who created only few mind maps is around 3% to 4%. Users who created more than 100 mind maps have a CTR of 5.26% on average. However, these numbers include all users, even those who started Docear only a few times and then decided not to use Docear again. Hence, most users (60.02%) created less than four mind maps.



	1	[2;3]	[4;5]	6-10	11-20	21-50	51-100	>100
Users	2.81%	7.62%	11.13%	25.58%	28.28%	18.66%	4.51%	1.40%
CTR	3.52%	2.89%	3.80%	4.03%	4.83%	5.16%	5.82%	5.26%

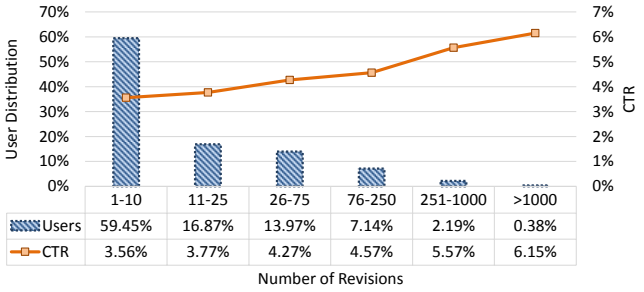
Number of Mind Maps

Figure 16: Number of Mind-Maps (Users Started Docear on 20 or more days)

Of those users, who started Docear on at least 20 days, most users created between six and 20 mind-maps (53.86%) and the correlation between mind maps and a CTR becomes clearer (Figure 16). Whiles users with less than six mind maps had CTR of less than 4% on average, users with more than 20 mind maps had CTR of more than 5% on average. Users who started Docear on at least 20 days and created between 51 and 100 mind maps, had the highest CTR of 5.82% on average.

CTR also correlates with the number of mind map revisions. For users who started Docear on at least 20 different days, more revisions led to a higher average CTR (Figure 17). Again, most users (59.45%) who started Docear on at least 20 days had 10 or less revisions of their mind maps. These "occasional" users achieved average CTR of only 3.56%. In contrast, users with more than 1,000 revisions are seldom (0.38%) but achieve the highest average CTR (6.15%).

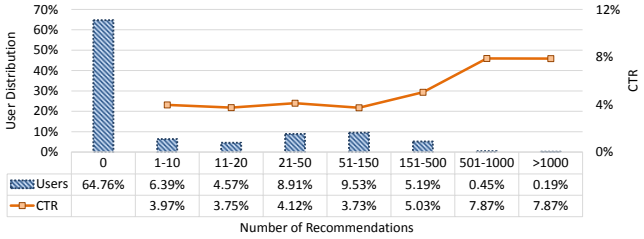
The average CTR for users who received 150 or less recommendations is around 4% (Figure 18). However, users who received more than 150 recommendation, have a significantly higher CTR on average (up to around 8%). Automatic recommendations are delivered only every five days of using Docear. Thus, the probability is high, that users, who have received many recommendations, actively requested most of them. We suspect users who request recommendations to find them generally more valuable, which results in a higher CTR. We also believe that users are not at all times interested in recommendations. Thus, they are probably more likely to click documents they actively request, than documents they automatically receive at a potentially unfitting time.



	1-10	11-25	26-75	76-250	251-1000	>1000
Users	59.45%	16.87%	13.97%	7.14%	2.19%	0.38%
CTR	3.56%	3.77%	4.27%	4.57%	5.57%	6.15%

Number of Revisions

Figure 17: Number of Revisions (Users Started Docear on 20 or more days)

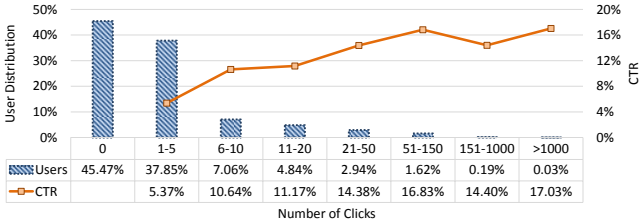


	0	1-10	11-20	21-50	51-150	151-500	501-1000	>1000
Users	64.76%	6.39%	4.57%	8.91%	9.53%	5.19%	0.45%	0.19%
CTR		3.97%	3.75%	4.12%	3.73%	5.03%	7.87%	7.87%

Number of Recommendations

Figure 18: Number of Recommendations received

Of those users who received at least one set of recommendations, the largest fraction (45.47%) did not click on any of them (Figure 19). The second largest fraction (37.85%) clicked between one and five recommendations. Only 4.78% of all users clicked recommendations more than 20 times. The more often a user has clicked recommendations, the higher the CTR tends to become.



	0	1-5	6-10	11-20	21-50	51-150	151-1000	>1000
Users	45.47%	37.85%	7.06%	4.84%	2.94%	1.62%	0.19%	0.03%
CTR		5.37%	10.64%	11.17%	14.38%	16.83%	14.40%	17.03%

Number of Clicks

Figure 19: Number of clicks

4. SUMMARY & DISCUSSION

In our paper we made two contributions.

First, we showed that user characteristics affect the usage and effectiveness of research paper recommender systems. In our scenario, i.e. the reference management software Docear and its recommender system, male users had higher average CTR than female users. In addition, male users used the recommender system more frequently and more intensively. The age of users also had an impact: older users achieved higher CTR than younger users on average. Not only demographics but also usage intensity affected CTR. The more intensive users had used Docear (e.g. more mind maps or mind map revisions they had created), the higher CTR became on average.

For most researchers it is probably not surprising that CTR differed for different user groups. However, to the best of our knowledge, for

recommender systems we are first to empirically show the differences in such a level of detail. In addition, we are quite certain that at least in the domain of *research paper* recommender systems, there is no comparable research. In a recent literature survey, we reviewed more than 200 papers on about 80 research paper recommender approaches, and none of them provided information about differences in effectiveness for different user groups [5].

Our results also show the importance of reporting user characteristics in research papers. If such information is missing, the readers of the articles cannot estimate whether the user populations are similar to the populations of their own systems, and hence how effective the evaluated recommendation approach would perform in their own system. For instance, if a recommendation approach achieves a CTR of 5% with a primarily male user population, the approach might achieve a significantly different CTR in an evaluation with primarily female users. Similarly, an approach evaluated with mainly new users will probably achieve significantly different CTR as if evaluated with power-users who are using the system for a longer time.

The second contribution of this paper is to provide detailed information on Docear's users. This information allows to better understand the context of our previously conducted evaluations and to assess whether Docear's recommendation scenario is similar to another scenario.

Our research has a few limitations. Our demographic information is currently limited to gender and age, but other demographics such as nationality and profession probably also have an impact and hence should also be reported in research papers. However, currently, Docear's users cannot provide such information during the registration process. In upcoming versions, we will adjust Docear to allow users to provide more information about themselves. Beside demographic data, other characteristics of a recommender system, (e.g. the quality of the data sets, the user interface or the intentions behind the recommender system) may influence a recommender system's effectiveness, and we have not yet researched these aspects.

Moreover, our research only considered whether specific user groups are more likely to click recommendations than others. Another interesting question is whether different recommendation approaches fit different user groups best. While, for instance, a specific recommendation approach might lead to the highest CTR for new users, another one might perform better for power-users. Finally, our analysis focuses on a unique scenario, i.e. research paper recommendations based on mind-maps. In the future, it should be studied how demographic data influence the usage and effectiveness of recommender systems in other scenarios.

5. REFERENCES

[1] Beel, J. et al. 2013. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (2013), 7–14.

[2] Beel, J. et al. 2011. Docear: An academic literature suite for searching, organizing and creating academic literature. *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (2011), 465–466.

[3] Beel, J. et al. 2013. Introducing Docear's research paper recommender system. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (2013), 459–460.

[4] Beel, J. et al. 2013. Persistence in Recommender Systems: Giving the Same Recommendations to the Same Users Multiple Times. *Research and Advanced Technology for Digital Libraries*. Springer. 386–390.

[5] Beel, J. et al. 2013. Research paper recommender system evaluation: a quantitative literature survey. *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (2013), 15–22.

[6] Beel, J. et al. 2014. The Architecture and Datasets of Docear's Research Paper Recommender System. *Proceedings of the 3rd International Workshop on Mining Scientific Publications (WOSP 2014) at the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014)* (2014).

[7] Beel, J. et al. 2013. The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. *Research and Advanced Technology for Digital Libraries*. Springer. 396–400.

[8] Beel, J. et al. 2014. Utilizing Mind-Maps for Information Retrieval and User Modelling. *Proceedings of the 22nd Conference on User Modelling, Adaption, and Personalization (UMAP)* (2014).

[9] Chittenden, L. and Rettie, R. 2003. An evaluation of e-mail marketing and factors affecting response. *Journal of Targeting, Measurement and Analysis for Marketing*. 11, 3 (2003), 203–217.

[10] Herlocker, J.L. et al. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*. 22, 1 (2004), 5–53.

[11] Krulwich, B. 1997. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI magazine*. 18, 2 (1997), 37.

[12] Pazzani, M.J. 1999. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*. 13, 5-6 (1999), 393–408.

[13] Said, A. et al. 2011. A comparison of how demographic data affects recommendation. *Adjoint Proc. of the 19th international conference on User modeling, adaption, and personalization* (2011).

[14] Uitdenbogerd, A.L. and Schyndel, R.G. van 2002. A Review of Factors Affecting Music Recommender Success. *ISMIR* (2002), 204–208.

[15] Weber, I. and Castillo, C. 2010. The demographics of web search. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), 523–530.